DOCUMENT RESUME

ED 091 434 TM 003 648

AUTHOR Kifer, Edward; Bramble, William
TITLE The Calibration of a Criterion-Referenced Test.
PUB DATE [Apr 74]
NOTE 23p.; Paper presented at the Annual Meeting of the American Educational Research Association (59th, Chicago, Illinois, April 1974)

EDRS PRICE MF-$0.75 HC-$1.50 PLUS POSTAGE
DESCRIPTORS Academic Ability; Achievement Tests; *Criterion Referenced Tests; Decision Making; Item Analysis; *Measurement Techniques; Prediction; Probability; Scores; Standard Error of Measurement; Test Construction; *Test Interpretation
IDENTIFIERS Rasch Model; *Scaling Techniques

ABSTRACT

A latent trait model, the Rasch, was fitted to a criterion-referenced test. Approximately 90 percent of the items fit the model. Those items which fit the model were then calibrated. Based on the item calibration, individual ability estimates and the standard errors of those estimates were calculated. Using the ability estimates, it was possible, given any criterion level, to compute the probability of a person exceeding the criterion. The calibration was done on the final examination of a performance-based introductory educational psychology course. (Author)

ED 091434

TM 003 438

U.S DEPARTMENT OF HEALTH
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION
THIS DOCUMENT HAS BEEN REPRO
DUCED EXACTLY AS RECEIVED FROM
THE PERSON OR ORGANIZATION ORIGIN
ATING IT. POINTS OF VIEW OR OPINIONS
STATED DO NOT NECESSARILY REPRE
SENT OFFICIAL NATIONAL INSTITUTE OF
EDUCATION POSITION OR POLICY

THE CALIBRATION OF A CRITERION-REFERENCED TEST

Edward Kifer

and

William Bramble

University of Kentucky

A Paper Prepared for the Annual AERA Meeting - Chicago, Illinois

April 15, 1974

THE CALIBRATION OF A CRITERION-REFERENCED TEST

I. Introduction

With increasing frequency the psychometric properties and problems of criterion-referenced tests are appearing in the research literature. Much of the discussion of these tests focuses on the extent to which classical test theory is an appropriate perspective from which to view criterion-referenced measurements. Where it is assumed, apparently, that classical test theory is a valuable approach for the estimation and interpretation of scores within a normative rubric, investigators increasingly question its efficacy for criterion-related interpretation of scores.

There is no consensus about what a criterion-referenced test is (See Popham and Husak, 1969; Glaser and Nitko, 1971; Kriewall, 1969 for somewhat different definitions). We think, however, that the crucial difference between criterion-referenced measurements and normative based testing is the way that scores are interpreted. The question is whether the test is designed for the purpose of comparing a score to some criterion or standard or whether it is designed to make comparisons between individuals. From what we can see, proponents of criterion-referenced measurement wish, above all, to provide an interpretation for a score, not in terms of other persons' scores but in terms of an absolute and arbitrary criterion.

If the crux of the difference is one of score interpretation, many important questions are raised by the use of criterion-referenced measurements: 1) Assuming that a standard or criterion has been chosen, how generalizable is it? That is, is the standard appropriate for only one particular group or is it an expectation for a general category of clients. Assuming the latter,

are the test statistics or item statistics used in judging the psychometric properties of the test the appropriate ones? 2) What is the relationship between the test items and test scores and those standards? Should there be a correspondence between what the items measure and the criteria used to certify that a score represents ability at or above the standard? 3) Given that the scores are compared to a standard, with how much precision can one state whether a particular score represents attainment or above the standard? That is, how good is the decision which classifies students as having either attained or not attained that standard?

Though these questions do not exhaust the domain of those which can be asked about criterion-referenced tests, they include, we hope, some of the more important ones concerned with interpretation of scores. It is with the these questions in mind that we chose to calibrate a criterion-referenced test using the Rasch model (Wright and Panchepakesan, 1968), one of a class of latent trait models. We have few illusions about our ability to answer definitively any or all of the questions but we do hope that our approach to the problems will provoke discussions of the issues we raise.

II. The Procedures

The data for this investigation are item responses on an 84 item final examination from an introductory educational psychology course at the University of Kentucky. The course is considered experimental, is organized around a series of learning tasks and the final examination contains items which are keyed to performance objectives. The objectives are, of course, the desired outcomes of the series of learning tasks. We think this final examination meets the minimum criteria for being criterion-related. That is, each of the items is keyed to a performance objective and the measurement goal is to compare students to an arbitrary criterion.

The subjects are 201 University of Kentucky undergraduates enrolled in the College of Education. All education students are required to take the course.

The items for the test were fitted to the Rasch model, a special form of the logistic response model. The estimation procedure, using maximum likelihood estimation, produces simultaneously estimates of item difficulty and subject ability. The process by which the test is calibrated is based on the procedures outlined by Wright and Panchepakesan (1968) and includes the following steps:

1) All of the items are calibrated and ability estimates are obtained based on all responses.

2) The items which are judged as fitting the model (Chi Square $\leq .05$ for rejection) are then refitted resulting in new item estimates and new ability estimates based only on items which fit the model.

When the items are calibrated, the measurement consequences are as follows (Wright and Panchepakesan, 1968):

1) Each of the items measures a latent trait and each is calibrated according to its difficulty.

2) Any subset of the items can be used to measure the same ability.

3) The estimates of item difficulty and person ability are invariant from sample to sample.

4) The ability estimates are on an interval scale.

5) The scale or metric of the abilities is defined regardless of the subjects who take the test. (That is, there is no necessity to interpret scores within a normative framework.)

6) Associated with each ability level is a unique standard error of measurement.

With the items calibrated and the abilities estimated, we proceed with a determination of whether a particular score exceeds an arbitrary criterion. The method we use makes very strong assumptions about the scores: Given any criterion we assume that the estimate of latent ability at the criterion is an estimate of the "true" ability at that point. The standard

error of measurement associated with that ability level is assumed to be an estimate of the observed score standard deviation around the "true" criterion. Based on those assumptions, it is possible to ask the question of the probability that any observed score comes from that particular distribution. Although the choice of sampling distributions for our estimates is arbitrary, because maximum likelihood estimates are asymtotically normal, we choose the normal distribution. For more conservative limits, one could use Tchebysheffs Inequality.

The next and last step is to ask the question of how good our decision is. It is possible to make two types of mistakes: 1) Falsely classify persons as meeting the criterion when they do not (false positive); 2) Falsely classify persons as not meeting the criterion when they actually do (false negative). If the two types of decisions are equally costly, then the probabilities of meeting the criterion that we assigned in the step above is the basis for our decision. If the costs of mis-classifying are not equal, then the ratio of those costs will affect our decision rule, or effectively switch the location of the distribution of interest. If it is 10 times as costly to mis-classify a person as having not met criterion when he does as it is to make the opposite mistake, then we will classify people as having met the criterion even though the probability, that they did, is rather small. By introducing a loss function, decisions we make are, therefore, dependent on not only the probability statements but also on the consequences of making particular kinds of decisions (See Hambleton & Novick, 1973 for a more rigorous discussion of the loss function).

III. The Results

Table 1 represents the results of the first step of the procedure for fitting a latent trait model to test items - that is, fitting all items

and all responses to the model. If one looks only at the probability that an item fits the model, then 16 items would be eliminated from the pool. (Those items are marked with a star.) This represents a problem for us since those 16 items are about 19% of the total. We will deal later in the paper with the question of whether it is reasonable to throw out such a large number of items. In practice, the items could be eliminated, re-written to improve their quality, or included in the test because they are good items which measure important performances. For this project we merely eliminated the 16 items and re-calibrated the ones which remained. (See Table 2)

Using the 68 items which we re-calibrated we can assign a probability that each person reached the criterion. As an example we take 80 per cent correct as the criterion. (See Table 3 for ability estimates and their standard errors.) The corresponding ability estimate is 1.59 with a standard error of .31. Using the standard error as an estimate of the standard deviation of the distribution of observed scores around a true score at the criterion, and using the normal distribution of errors, it is possible to give the probability that any observed score comes from a distribution with 1.59 as its mean and .31 as its standard deviation. For example, an ability of 2.17 has approximately .03 probability of being part of the distribution. An ability of .966 has approximately .02 probability of coming from that distribution. Since the first ability is above the criterion we would decide that the person with that score has exceeded the criterion. On the other hand a person with the second score would be assumed not to have exceeded the criterion because the score is so far below it. (One should note that the estimated abilities at the above levels are easy to deal with. As scores move closer to the criterion, the decision becomes less straight forward.)

Based on this criterion level and its standard error, it is possible to compute a probability for each estimated ability level. And, as is evident, one can use this same process regardless of the level chosen for the criterion.

All other things being equal, we would select those persons whose ability estimates given the distribution, have probability .5 or better of meeting the criterion. The long run consequences of those decisions would be to insure that those we select have true scores above the criterion while those not so selected do not. If all things are not equal, that is, the cost of the decision varies by whether it is false positive or false negative, we would then add a loss function.

As already mentioned, decisions about estimated abilities exceeding or not exceeding the criterion ability are straight forward where the estimated abilities are sufficiently different from the criterion. However, in the case of estimated scores close to the criterion the decisions are more difficult, especially if one considers the losses associated with the two false decisions which can be made. For example, consider the estimated ability of 1.488 corresponding to a score of 53 on the test. The standard error of this estimate is .311. Considering a normal distribution with mean at 1.488 and standard deviation .311 the difference between the score and the criterion is .318 standard deviations. One could estimate that the probability that a particular student with estimated ability $\hat{\tau}_j = 1.488$ has a true ability (τ_j) that exceeds the criterion is .37 and the probability that his true ability does not exceed the criterion is .63. That is, $P(\tau_j > c | \hat{\tau}_j = 1.488) = .37$ and $P(\tau_j < c | \hat{\tau}_j = 1.488) = .63$. The appropriate decision to make here depends on the loss associated with false positives and false negatives. It might be that the cost of retraining a student is

far greater (let's say by a ratio of 4:1) than the cost associated with falsely certifying that the student has met the criterion. Thus the loss matrix has the form:

		True State: meets criterion	True State: doesn't meet criterion
Decision	meets criterion	0	1
	doesn't meet criterion	4	0

The expected loss for the decision "meets criterion" is:

$$E(l_p) = l_{12p}(\tau < c \mid \hat{\tau} = 1.488)$$

$$= (1)\ (.63) = .63$$

The expected loss for the decision "doesn't meet criterion" is:

$$E(l_f) = l_{21p}(\tau > c \mid \hat{\tau} = 1.488)$$

$$= 4\ (.37) = 1.48$$

Thus the loss associated with false negatives is clearly greater than that for false positives and this would lead us to classify individuals with $\hat{\tau} = 1.488$ as having met the criterion.

Of course different decisions can be made with different loss values. The optimum action, however, for any score group would be to decide that a student has met the criterion if $\dfrac{P(\tau_k > c \mid \hat{\tau}_k)}{P(\tau_k < c \mid \hat{\tau}_k)} > \dfrac{l_p}{l_f}$

The ratio l_p/l_f in the problem being discussed is ½. For a student with estimated ability 1.488 the ratio of $P(\tau_k > c \mid \hat{\tau}_k)/P(\tau_k < c \mid \hat{\tau}_k)$ has the value $\frac{.37}{.63}$. Thus the optimum action is to classify the student as meeting the criterion. Notice that if other losses are considered such that $l_p/l_f > \frac{.37}{.63}$ the opposite decision would be made.

IV. Discussion

Before discussing the advantages of our procedure, it is necessary to discuss some possible limitations. The first, and major limitation, may be the problem of arriving at a decison about what to do with items which measure important performances but do not fit a latent trait. In the case of this study, 19% of the items presented this dilemma. While we are aware that difficulties like these inevitably can be explained away on a post hoc basis, we prefer only to explain some of our findings with the appropriate reservations.

Although there can be many reasons that an item does not fit a latent trait, there are three main ones for the Rasch model: 1) The item measures a different trait; 2) The item is poorly written and does not measure the desired performance; 3) The item measures the trait but does not fit the model because of the model's restrictive assumptions (i. e., that all the discrimination parameters are equal to one). When we looked at the 16 items which did not fit the trait, we found things which indicated that some of the items should not be included in the test. We estimated that 40% of these 16 items contained errors which could be corrected by re-writing. The items either presented ambiguous problems in the stem or response alternatives which made more than one response seem correct.

Another possibility with the 16 items is to re-calibrate them to see if they measure a separate trait. We did this, and although the 16 items did fit a trait (See Table 4), the imprecision of the estimates of the item parameters was noteably. Ability estimates based on these items were also imprecise so we suspect that if a trait is being measured by these items it is being measured so poorly that little information can be gained from them.

We had no opportunity to re-write the items or generate new items to

measure the performances of interest, so we do not know for sure that the fit of the test could be improved by better item writing. Our best guess, based on a look at the items which did not fit and an attempt to re-calibrate those same items, is that the fit of the model could be greatly improved. Without stretching the point too far, we believe that the Rasch model, being fitted to a criterion-referenced test, could fit 95% of the items if they were well written and if the domain that is sampled contains only one trait.

For other tests in other situations we think it is both desirable and feasible to fit more than one latent trait. There is every reason to believe that criterion-referenced tests can be related to more than one trait. If this is the case, we would advocate an **a priori** determination of which items measure which traits and then calibrating those item sets separately. By so doing, one retains the power of the model to give the test maker an indication of which items may be of limited utility.

The above should not be construed to mean that we think items can be eliminated from a criterion-referenced test simply because they do not fit a latent trait model. We can think of many situations in which it would be perfectly reasonable to keep items which apparently measure performances of interest and appear to be good items, regardless of whether they fit a trait. What we are advocating, however, is that the test maker should hypothesize which items measure which trait, calibrate the items, and look closely at items which do not fit the model to see if the items are good ones. If they are not good ones, re-write them; if they are good ones, keep them.

A second limitation of this procedure is the size of the samples needed to generate stable parameter estimates. Wright (personal communication) suggests a group of at least 200 respondees. This suggests that our procedure is a practical technique only for test makers who have at their disposal a large

number of subjects and wish to generate criterion-referenced measures for a for a large population. As will be discussed later, however, if a large number of items generated from a domain of knowledge are calibrated, classroom teachers could benefit from our procedure.

If one chooses to calibrate criterion-referenced measures with a latent trait model, important consequences result. Before describing those consequences, however, it is necessary to make an assumption explicit: We believe that in practice it is impractical, and perhaps impossible, to separate criterion-referenced measurements from a normative rubric. In a fundamental sense the test maker imposes a normative framework on his criterion-referenced test when he defines a domain of items or decides which testing objectives should be excluded or included in the development stage of the test. (See Millman, 1973 for a brief description of different approaches to the problem.) In defining a domain, the person constructing a criterion-referenced test can be as specific or general as is desirable. The only way to validate such a procedure is to resort to a group of experts who concern themselves with the appropriateness of the test for the situation in which it is to be used. In the end, a normative rubric, is, for all practical purposes, being established.

The second way in which criterion-referenced tests become normative is when persons set a criterion level. Logically, a criterion-referenced test should be "graded" dichotomously - a person either gets all of the items correct or he does not. By establishing an 80% criterion level, for example, the test maker is saying implicitly that it does not matter which 80% of the items the person gets correct. But yet, some items may measure more important performances than others, and in terms of the way the test was constructed, it may be far more important to know what items are missed rather than what percentage is correct. By establishing a criterion level,

the test constructer may be doing one of two things: 1) saying that he will make a decision about exceeding a criterion regardless of the content of the items; or, 2) saying he will make his decision about the criterion based on the sample of students who took the test. In either case, a kind of normative statement is being made.

A calibration of items will eliminate the problems stated above. A property of the Rasch model is that it leads to estimates of item parameters which are independent of the sample of persons who took the test. Also, estimates of ability are on a similar metric regardless of the items which are attempted. (Wright, 1968 calls this sample free test calibration and person measurement.) The result of the calibration of a test is the possibility of generating scores which do not need a normative framework for score interpretation, and do not depend either on the scores of other persons who took the test or the sample of items which were on the test. We, therefore, can make comparisons between scores and a standard or criterion which means the same thing regardless of the persons who take the test or the items included in the test. For example, if a person with a score of 2.0 is considered above the criterion, any person with that score, based on any sample of the calibrated items, can be assigned a probability of having met the criterion. Classical test theory will not lead to such statements because a criterion of X per cent in classical test theory cannot be separated from either the group who took the test or the items which were used to measure the ability. A person who applied classical test theory to a criterion-referenced test may be in the less than enviable position of having criterion-referenced measurements mean something only within a very small and specific group. To defend the use of classical test theory in this case, is either to make statements about the sample who took the test or be required to defend, perhaps blindly, the criterion or domain that was chosen.

A more important consequence of calibrated items, however, is in regard to the standard error of measurement. In classical test theory, this measure of precision is related generally to the reliability of the test. Because the reliability of criterion-referenced tests is often low, various reliability coefficients have been generated for such tests. That approach and its potential weaknesses are well-documented (Livingston, 1972 (a); 1972 (b); Harris, 1972; Shavelson et.al., 1972). The latent trait model eliminates the question of reliability because 1) it leads to an estimate of the standard error of measurement for each level of ability (and, based on our assumptions, a different error at each possible criterion level); and, 2) given a sufficient pool of items, it is possible to estimate a person's score to the degree of precision that the tester deems necessary.

Because each ability level has an unique standard error associated with it, the probability of a person exceeding the criterion depends both on the level of the criterion and the errors of measurement at that level. Standard errors of measurement based on a reliability coefficient do not generally have that property. Hence, precision and accuracy of the probability statements can be increased by using a latent trait model. (Property 1 above is useful not only for the interpretation of scores but also for the construction of the test.) If the test constructor knows the criterion level a priori, the fact that each ability estimate has associated with it a unique standard error of measurement, leads to a guide for constructing items. That is, items should be written so that their levels of difficulty are as close to the criterion as possible. In that way the test constructor will assure himself of having a relatively small standard error of measurement at the criterion. That, of course, will lead to a narrower distribution of observed scores around a "true" criterion. In addition, if an item pool contains a large proportion of items with difficulties at the criterion, additional information about

subjects (assuming that they have not taken all items) could be gained by administering more and more items at the criterion level.

Because it is possible, given a pool of calibrated items, to estimate person abilities to a given level of precision, Wood (1973) has advocated the use of latent trait model for "Response-Contingent Testing." Such a procedure could be modified or adapted to criterion-referenced testing and be of enormous use to both test makers and classroom teachers. Given a large pool of calibrated items from a criterion-referenced test, the test maker could construct various forms of the tests for different purposes. For example, one form could be constructed which sampled a wide range of behaviors; another form could be composed of mainly items at the criterion level so teachers could make precise determinations of whether persons exceeded the criterion; another form could measure just a few of the objectives. With an item pool which is criterion-referenced and measures a latent trait, classroom teachers could potentially have access to powerful and important diagnostic tools.

V. Conclusion

The Rasch model, and other latent trait models, are potentially useful for the calibration of criterion-referenced tests. If the items fit a latent trait there are important consequences in terms of the interpretation of scores and the measurement of abilities. With a large pool of calibrated items, the tester can use any sub-set of them to estimate ability on the trait. This makes the comparison between the student's ability and the criterion a meaningful one without resorting to either a normative rubric or a defense of the criterion and the domain from which the items are selected. Given a criterion, it is possible to say, without regard to a sample of persons or the sample of items what the probability of a person meeting the criterion is. In addition, given a large pool of items it is possible to estimate person

ability to practically any degree of accuracy. This insures a more precise determination of the probability that the person has exceeded the criterion.

We are convinced then that the calibration of a criterion-referenced test leads to more consistency between the purposes of such testing and the interpretation of the scores. It leads to more generalizability about the meaning of the scores and more precision concerning the extent to which a score represents passing a criterion. Despite some limitations of this procedure it has great potential in its application.

ITEM PARAMETERS

Table 1

ITEM LABEL	ITEM DIFFICULTY	STANDARD ERROR	MEAN SQUARE	PROBABILITY
1	-1.871	0.329	0.783	0.848
2	-1.169	0.248	0.884	0.690
3	-0.671	0.208	0.984	0.502
4	2.104	0.161	1.779	0.001
5	-0.899	0.225	0.676	0.951
6	1.907	0.157	2.058	0.000
7	0.377	0.162	1.209	0.162
8	-2.589	0.456	0.862	0.728
9	-0.629	0.206	1.131	0.255
10	-0.671	0.208	1.969	0.000
11	-1.434	0.274	0.687	0.944
12	-1.590	0.292	0.796	0.831
13	1.132	0.152	1.405	0.039
14	-0.850	0.221	1.156	0.221
15	[illegible]	0.167	1.002	0.468
16	[illegible]	0.153	1.474	0.022
17	[illegible]	0.161	0.948	0.570
18	[illegible]	0.152	0.783	0.849
19	1.244	0.151	1.518	0.015
20	-1.001	0.233	0.895	0.669
21	1.536	0.152	1.380	0.048
22	0.164	0.168	1.098	0.303
23	-0.162	0.181	0.785	0.846
24	-0.194	0.182	0.962	0.543
25	0.402	0.162	1.431	0.032
26	0.272	0.165	0.705	0.930
27	0.478	0.160	0.948	0.570
28	-0.549	0.201	2.353	0.000
29	-1.362	0.267	1.439	0.030
30	-0.130	0.179	1.104	0.294
31	-0.363	0.190	1.004	0.463
32	0.696	0.156	1.022	0.431
33	-0.671	0.208	1.148	0.232
34	0.601	0.157	1.000	0.472
35	0.325	0.164	0.958	0.551
36	-0.850	0.221	1.165	0.210
37	0.325	0.164	0.882	0.693
38	-0.671	0.208	0.778	0.855
39	-0.472	0.196	0.771	0.863
40	-0.130	0.179	1.042	0.395
41	-1.001	0.233	0.700	0.934
42	0.453	0.160	0.911	0.639
43	0.860	0.154	1.100	0.300
44	-2.589	0.456	0.965	0.538
45	-0.629	0.206	0.936	0.592

Table 1 (cont.)

ITEM LABEL	ITEM DIFFICULTY	STANDARD ERROR	MEAN SQUARE	PROBABILITY
46	0.219	0.167	1.532	0.013
47	1.650	0.153	2.113	0.000
48	0.022	0.173	0.885	0.688
49	0.673	0.156	1.063	0.360
50	1.513	0.152	1.629	0.005
51	-1.231	0.253	0.688	0.943
52	-0.435	0.194	1.453	0.026
53	-1.001	0.233	0.955	0.557
54	0.503	0.159	1.036	0.406
55	0.649	0.157	1.040	0.398
56	0.051	0.172	0.962	0.544
57	-0.803	0.217	0.906	0.649
58	0.022	0.173	1.031	0.414
59	-0.130	179	0.951	0.565
60	-0.629	06	1.190	0.182
61	1.513	52	0.609	0.981
62	-0.293	187	1.534	0.013
63	-0.435	0.194	0.848	0.752
64	-1.231	0.253	1.111	0.284
65	1.445	0.152	0.738	0.900
66	0.377	0.162	1.193	0.178
67	0.860	0.154	1.519	0.015
68	0.428	0.161	0.957	0.554
69	-0.259	0.185	1.067	0.353
70	-1.509	0.283	1.401	0.040
71	-0.671	0.208	1.354	0.059
72	0.022	0.173	1.011	0.451
73	0.929	9.153	1.119	0.272
74	0.576	0.158	1.247	0.126
75	1.109	0.152	1.326	0.072
76	0.246	0.166	1.221	0.149
77	1.042	0.152	0.951	0.565
78	1.931	0.157	1.180	0.192
79	-1.362	0.267	0.935	0.595
80	-0.194	0.182	0.977	0.514
81	0.576	0.158	1.109	0.287
82	0.325	0.164	1.017	0.440
83	1.109	0.152	0.966	0.536
84	0.051	0.172	0.672	0.953

ITEM PARAMETERS

Table 2

ITEM LABEL	ITEM DIFFICULTY	STANDARD ERROR	MEAN SQUARE	PROBABILITY
1	-1.786	0.331	0.669	0.941
2	-1.079	0.250	1.103	0.304
3	-0.573	0.211	0.624	0.966
4	-0.805	0.227	0.718	0.902
5	0.501	0.165	1.277	0.118
6	-2.505	0.458	0.756	0.861
7	-0.531	0.208	1.420	0.045
8	-1.346	0.276	0.856	0.721
9	-1.503	0.294	0.896	0.654
10	-0.756	0.223	1.094	0.318
11	0.310	0.171	0.711	0.908
12	0.554	0.164	0.866	0.703
13	1.169	0.156	0.772	0.842
14	-0.908	0.235	0.978	0.508
15	0.281	0.171	1.994	0.000
16	-0.053	0.183	0.742	0.878
17	-0.086	0.185	0.949	0.559
18	0.393	0.168	0.653	0.951
19	0.606	0.163	0.995	0.478
20	-0.021	0.182	0.649	0.953
21	-0.259	0.193	0.809	0.792
22	0.832	0.159	1.289	0.109
23	-0.856	0.211	1.073	0.349
24	0.733	0.161	1.084	0.333
25	0.448	0.167	0.880	0.681
26	-0.756	0.223	0.930	0.593
27	0.448	0.167	0.551	0.989
28	-0.573	0.211	0.867	0.702
29	-0.370	0.199	1.181	0.205
30	-0.021	0.182	0.747	0.872
31	-0.908	0.235	0.964	0.532
32	0.580	0.164	1.000	0.469
33	1.002	0.157	1.578	0.013
34	-2.505	0.458	0.911	0.627
35	-0.531	0.208	0.931	0.591
36	0.135	0.176	0.685	0.930
37	0.808	0.160	0.935	0.584
38	-1.141	0.255	0.741	0.878
39	-0.908	0.235	0.709	0.910
40	0.632	0.163	1.226	0.160
41	0.783	0.160	1.154	0.237
42	0.165	0.175	0.927	0.599
43	-0.708	0.220	0.819	0.777
44	0.135	0.176	0.837	0.750
45	-0.021	0.182	1.056	0.377

Table 2 (con't)

ITEM LABEL	ITEM DIFFICULTY	STANDARD ERROR	MEAN SQUARE	PROBABILITY
46	-0.531	0.208	0.984	0.497
47	1.686	0.156	0.936	0.582
48	-0.333	0.197	0.890	0.664
49	-1.141	0.255	0.727	0.893
50	1.615	0.156	1.302	0.101
51	0.501	0.165	0.800	0.805
52	1.002	0.157	1.278	0.117
53	0.554	0.164	1.063	0.366
54	-0.153	0.188	1.366	0.066
55	-0.573	0.211	0.868	0.700
56	0.135	0.176	1.205	0.180
57	1.074	0.157	0.776	0.837
58	0.708	0.161	1.582	0.013
59	1.263	0.156	0.912	0.625
60	0.366	0.169	1.411	0.048
61	1.192	0.156	1.122	0.279
62	2.126	0.162	1.257	0.133
63	-1.274	0.269	0.962	0.536
64	-0.086	0.185	0.748	0.870
65	0.708	0.161	0.794	0.814
66	0.448	0.167	0.796	0.811
67	1.263	0.156	1.152	0.240
68	0.165	0.175	0.876	0.686

ABILITY ESTIMATES

Table 3

SCORE GROUP	GROUP ABILITY	STANDARD ERROR
1	-4.689	1.021
2	-3.955	0.735
3	-3.510	0.610
4	-3.184	0.536
5	-2.923	0.487
6	-2.705	0.450
7	-2.515	0.422
8	-2.346	0.400
9	-2.194	0.381
10	-2.054	0.366
11	-1.925	0.353
12	-1.804	0.342
13	-1.690	0.332
14	-1.583	0.324
15	-1.481	0.316
16	-1.383	0.310
17	-1.289	0.304
18	-1.198	0.299
19	-1.110	0.294
20	-1.025	0.290
21	-0.942	0.286
22	-0.861	0.283
23	-0.782	0.280
24	-0.705	0.277
25	-0.629	0.275
26	-0.554	0.273
27	-0.480	0.271
28	-0.407	0.269
29	-0.335	0.268
30	-0.264	0.267
31	-0.193	0.266
32	-0.122	0.265
33	-0.052	0.265
34	0.018	0.265
35	0.088	0.265
36	0.159	0.265
37	0.229	0.265
38	0.299	0.266
39	0.370	0.267
40	0.441	0.268
41	0.513	0.269
42	0.586	0.270
43	0.660	0.272
44	0.734	0.274
45	0.810	0.277

Table 3 (con't)

SCORE GROUP	GROUP ABILITY	STANDARD ERROR
46	0.888	0.279
47	0.966	0.282
48	1.047	0.286
49	1.130	0.290
50	1.215	0.294
51	1.303	0.299
52	1.394	0.304
53	1.488	0.311
54	1.587	0.318
55	1.691	0.326
56	1.800	0.335
57	1.916	0.346
58	2.040	0.359
59	2.174	0.374
60	2.321	0.392
61	2.483	0.414
62	2.665	0.442
63	2.876	0.478
64	3.127	0.528
65	3.443	0.601
66	3.877	0.727
67	4.599	1.014

ITEM PARAMETERS

Table 4

ITEM LABEL	ITEM DIFFICULTY	STANDARD ERROR	MEAN SQUARE	PROBABILITY
1	1.410	0.160	0.528	0.871
2	1.236	0.156	1.936	0.036
3	-1.012	0.206	0.703	0.723
4	0.558	0.149	1.505	0.130
5	0.421	0.150	1.278	0.237
6	0.655	0.149	1.206	0.281
7	0.910	0.150	1.760	1.668
8	-0.075	0.159	0.917	0.516
9	-0.905	0.199	1.284	0.233
10	-1.623	0.265	0.464	0.914
11	-0.234	0.164	0.536	0.866
12	1.009	0.152	1.184	1.296
13	0.890	1.150	0.677	0.747
14	-0.805	0.192	1.285	0.232
15	-0.681	0.185	1.424	0.162
16	-1.754	0.281	0.910	0.523

BIBLIOGRAPHY

Glaser, R. & Nitko, A. J. Measurement in Learning and Instruction. In R. L. Thorndike (ed.) Educational Measurement. Washington: American Council on Education, 1971, pp. 625-670.

Hambleton, R. K & Novick, M. R. Toward an Integration of Theory and Method for Criterion-Referenced Tests. Journal of Educational Measurement. 1973, pp. 10, 3, 159-169.

Harris, C. W. An Interpretation of Livingston's Reliability Coefficient for Criterion-Referenced Tests. Journal of Educational Measurement. 1972, pp. 9, 27-29.

Kriewall, T. E. Applications of Information Theory and Acceptance Sampling Principles to the Management of Mathematics Instruction. Unpublished doctoral dissertation, University of Wisconsin, 1969.

Livingston, S. A. Criterion-Referenced Applications of Classical Test Theory. Journal of Educational Measurement. 1972, pp. 9, 13-26 (a).

Livingston, S. A. A Reply to Harris "An Interpretation of Livingston's Reliability Coefficient for Criterion-Referenced Tests". Journal of Education Measurement. 1972, pp. 9, 31 (b)

Lord, F. M. & Novick, M. R. Statistical Theories of Mental Test Scores. Reading, Massachusetts: Addison-Wesley, 1968.

Millman, J. Passing Scores and Test Lengths for Domain-Referenced Measures. Review of Educational Research. 1973, pp. 43, 2, 205-215.

Popham, W. J. & Husek, T. R. Implications of Criterion-Referenced Measurement. Journal of Educational Measurement. 1969, pp. 6, 1-9.

Shavelson, R. J., Block, J. H. and Ravitch, M. M. Criterion-Referenced Testing-Comment on Reliability. Journal of Educational Measurement. 1972, pp. 9, 2, 133-137.

Wood, R. Response-Contingent Testing. Review of Educational Research. 1973, pp. 43,4, 529-544.

Wright, B. D. Sample-Free Test Calibration and Person Measurement. In Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton, N. J.: Educational Testing Service, 1967.

Wright, B. D. & Panchapakesan N. A Procedure for Sample-Free Item Analysis. Educational and Psychological Measurement. 1969, pp. 29, 23-48.